Birla Institute of Technology & Science, Pilani

Work Integrated Learning Programmes Division
M.S/M Tech. Software Engineering at _____
First Semester _____
Comprehensive Exam (Regular)

Course Number	:	____ ZC425
Course Title	:	DATA MINING
Type of Exam	:	Open Book
Weightage	:	50 %
Duration	:	2.5 hours
Date of Exam	:	

No. of Pages : 3

No. of Questions : 6

Session : AN

---

1. Provide answers with very brief justification **(5 X 2 marks)**
   a. Calculate distance between the following clusters when you take a) Single-link approach b) Complete-link approach

      {2, 4, 10} and {20, 30, 311}

   b. Following table shows sales of a product over various days. Is there any evident pattern that can be mined without using any of data mining tasks?

Date	Sales (in Rs)	Date	Sales (in Rs)
1-1-2016	200	13-6-2016	200
15-1-2016	200	18-7-2016	200
15-12-2016	100	1-4-2016	500
14-12-2016	200	15-4-2016	600
1-7-2016	300	29-5-2016	600

   c. For the following vectors x and y, calculate the cosine similarity and euclidean distance measures:

      x =(4,4,4,4), y=(2,2,2,2)

   d. Consider the one-dimensional data set shown on the below table

X	0.6	3.2	4.5	4.6	4.9	5.2	5.6	5.8	7.1	9.5
Y	-	-	+	+	+	-	-	+	-	-

   Classify the data point x=5.0 according to its 3- and 9- nearest neighbors (Using majority Vote)

   e. A data analyst found that bread and cheese sell together often. Individually bread sells more often than cheese. Will you conclude that bread=>cheese or cheese=>bread?

2. Provide concise responses **(6 X 3 marks)**
   a. Given the following two objects with four binary attributes.

	Attribute1	**Attribute2**	**Attribute3**	**Attribute4**
Object1	1	1	0	0
Object2	1	0	1	0

a) What is the distance between the objects if variables are symmetric?
b) What is the distance between the objects if variables are asymmetric?
c) What is Jaccard coefficient for the objects?

b. Suppose a group of 12 sales price records has been sorted as follows:
5; 10; 11; 13; 15; 35; 50; 55; 72; 90; 204; 215:

Partition them into three bins by each of the following methods.
(a) equal-frequency partitioning
(b) equal-width partitioning
(c) clustering

c. How do Fraud Detection Systems make use of DM techniques?
d. Differentiate bagging and boosting techniques for enhancing classifier accuracy.
e. How is multimedia data mining different from business data mining?
f. A database has frequent 50-itemset. What are the minimum number of candidates generated by Apriori algorithm?

3. A database has five transactions. Let min sup = 60% and min conf = 80%. **(5+2 marks)**

TID	items bought
T100	Bread, Butter, Beans, Potato, Jam, Milk
T200	Bread, Butter, Shampoo, Potato, Jam, Milk
T300	Beans, Soap, Butter, Bread
T400	Beans, Onion, Apple, Butter, Milk
T500	Apple, Banana, Jam, Bread, Butter

(a) Find all frequent itemsets using FP-growth algorithm.

(b) List all of the strong association rules (with support s and confidence c) matching the following
buys(X; item1) ^ buys(X; item2) => buys(X; item3) [s; c]

4. The following table shows the house area(in sq meters) and house price(in lakh rupees) obtained for a metropolitan city. **(3+2 marks)**

Area	72	50	81	74	94	86
Price	84	63	77	78	90	75

(a) Assuming linear relationship, Use the method of least squares to get an equation for the prediction of price based on the area.
(b) Predict the price of a house with 80 sq. meter area.

5. Apply K-Means clustering algorithm for the set of data in the table below, where K value is 2 (i.e. No. of Clusters or Groups = 2). Each object has two numeric attributes A and B.

(5 marks)

Object	1	2	3	4	5	6	7	8	9	10	11
A	1	1.5	3	3.5	4	7	5.5	6	6.5	7	5
B	1	2	4	5	4.5	10	5	8.5	8	7.5	7

6. How does F-score help in quantifying cluster quality? Given clustering results of newspaper articles data set given below, compute F-score for cluster representing metro and financial news articles. (Column titles are actual classes of documents.) List any assumptions made.

(2+3 marks)

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Total
#1	1	1	5	11	4	676	698
#2	27	89	333	827	253	33	1562
#3	126	465	8	105	16	29	749
Total	154	555	346	943	273	738	3009

XXXXXX